BUSINESS ANALYTICS IN ACTION WITH RSTUDIO

Rishitha Fernando

Abstract

Business Analytics (BA) encompasses a set of methodologies, tools, and techniques designed to extract valuable insights from data, enabling businesses to make data-driven decisions, optimize processes, and identify new opportunities. This paper explores key BA methodologies, such as descriptive, predictive, and prescriptive analytics, and tools, including statistical software, data visualization platforms, and machine learning frameworks. Techniques like data mining, regression analysis, and hypothesis testing further aid in generating actionable insights. The value of business analytics is demonstrated through enhanced decision-making, operational efficiency, improved customer understanding, and competitive advantage. By applying these analytics frameworks to real-world business problems, organizations can not only address immediate challenges but also unlock long-term growth potential. Finally, a proposed solution is presented for a typical business problem, illustrating how the integration of appropriate BA methodologies, tools, and techniques can lead to optimized outcomes and drive business innovation.

Strategic Use of Data Science and Analytics in the Ministry of Industry and Commerce

Data science, analytics, and business intelligence (BI) empower the Ministry of Industry and Commerce in Sri Lanka to make data-driven decisions that enhance strategic planning, operational efficiency, and public service delivery.

- Senior-level officials leverage analytics for macroeconomic and industrial sector analysis, enabling evidence-based policy-making aligned with national goals.
- Mid-level officials use predictive tools to optimize supply chains and support SME growth through demand forecasting and market analysis.
- Junior-level officials apply feedback analysis and statistical tools like SPSS to improve day-to-day service delivery and responsiveness.

By integrating data analytics across all levels, the Ministry can drive sustainable development and improve decision-making effectiveness.

Tools, Techniques, and Methodologies Used

To analyze ministry-level data, a combination of statistical tools and methods was employed:

Tools:

- Excel – for data entry, filtering, and basic analysis.
- RStudio & R – for in-depth statistical analysis, visualizations, and regression modeling.

Techniques:

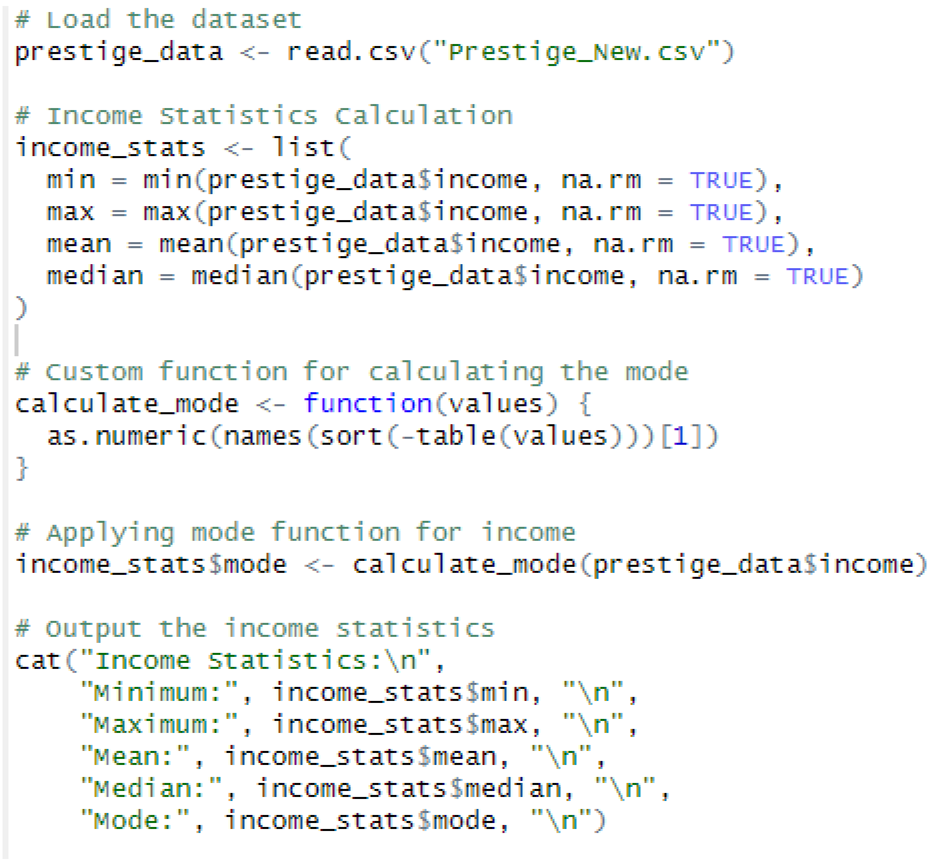
- Descriptive Statistics – to summarize data using mean, median, and standard deviation.
- Data Visualization – including histograms, scatterplots, and bell curves to uncover trends.
- Correlation & Regression Analysis – to explore relationships between variables.
- ANOVA Testing – to assess differences in performance across groups.

Methodologies:

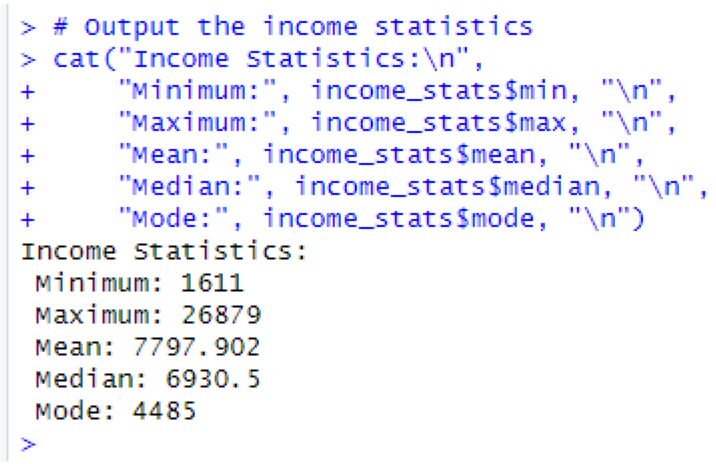
- Hypothesis Testing – to validate assumptions about relationships.
- Variance & Normality Testing – to ensure statistical reliability and guide model selection.

These approaches collectively enabled the extraction of actionable insights for better resource allocation and strategic planning within the Ministry.

Finding out Minimum, Maximum, Mean, Median & Mode of Income



```r
# Load the dataset
prestige_data <- read.csv("Prestige_New.csv")

# Income Statistics Calculation
income_stats <- list(
  min = min(prestige_data$income, na.rm = TRUE),
  max = max(prestige_data$income, na.rm = TRUE),
  mean = mean(prestige_data$income, na.rm = TRUE),
  median = median(prestige_data$income, na.rm = TRUE)
)

# Custom function for calculating the mode
calculate_mode <- function(values) {
  as.numeric(names(sort(-table(values)))[1])
}

# Applying mode function for income
income_stats$mode <- calculate_mode(prestige_data$income)

# Output the income statistics
cat("Income Statistics:\n",
    "Minimum:", income_stats$min, "\n",
    "Maximum:", income_stats$max, "\n",
    "Mean:", income_stats$mean, "\n",
    "Median:", income_stats$median, "\n",
    "Mode:", income_stats$mode, "\n")
```



```
> # Output the income statistics
> cat("Income Statistics:\n",
+     "Minimum:", income_stats$min, "\n",
+     "Maximum:", income_stats$max, "\n",
+     "Mean:", income_stats$mean, "\n",
+     "Median:", income_stats$median, "\n",
+     "Mode:", income_stats$mode, "\n")
Income Statistics:
 Minimum: 1611
 Maximum: 26879
 Mean: 7797.902
 Median: 6930.5
 Mode: 4485
>
```

The code loads the "prestige_New.csv" dataset and calculates key income statistics: minimum, maximum, mean, median, and mode. A custom `calculate mode` function determines the most frequent income value. These statistics are displayed using the `cat` function. Sample output includes:

- Minimum: 6111

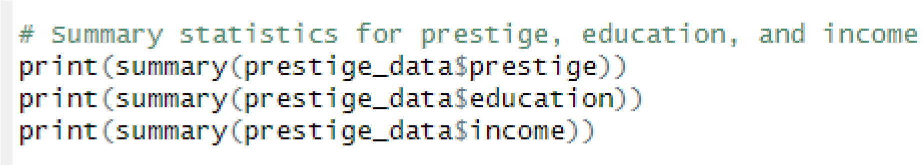
- Maximum: 25879

- Mean: 7797.902

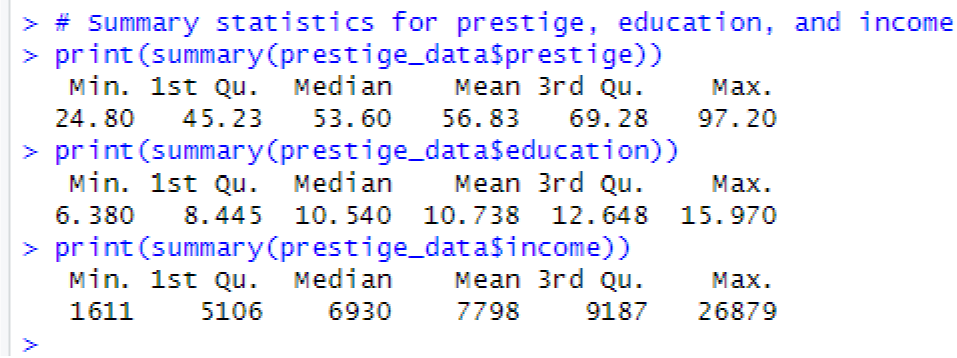
- Median: 6930.5

- Mode: 4485

Summary of statistics



```
# Summary statistics for prestige, education, and income
print(summary(prestige_data$prestige))
print(summary(prestige_data$education))
print(summary(prestige_data$income))
```



```
> # Summary statistics for prestige, education, and income
> print(summary(prestige_data$prestige))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  24.80   45.23   53.60   56.83   69.28   97.20 
> print(summary(prestige_data$education))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.380   8.445  10.540  10.738  12.648  15.970 
> print(summary(prestige_data$income))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1611    5106    6930    7798    9187   26879 
> 
```

The code calculates basic summary statistics for three variables: "prestige," "education," and "income" in our dataset. By calling the summary() function on each column, we get an overview of each variable's distribution, which helps us understand their central tendency, spread, and range. Here's how each component is interpreted and justified:

1. **Prestige:**

    - **Minimum (24.80):** This is the lowest prestige score in our dataset, showing the baseline for our prestige data.
    - **1st Quartile (45.23):** This value represents the point below which 25% of the prestige scores fall. It's useful for understanding the distribution's lower end.
    - **Median (53.60):** The median divides the prestige data into two equal halves, indicating a central point that isn't skewed by outliers.
    - **Mean (56.83):** The mean prestige score represents the average and provides a measure of central tendency across all values.
    - **3rd Quartile (69.28):** This value indicates that 75% of prestige scores fall below this point, showing us the upper end of the distribution.
    - **Maximum (97.20):** The maximum prestige score highlights the upper limit, providing a sense of the highest recorded prestige levels.

Justification*:* These values collectively help us understand how prestige is distributed, from minimum to maximum, with key percentiles that reveal its spread and central tendency. This summary helps us determine if prestige scores are clustered or widely dispersed.

2. **Education:**

    - **Minimum (6.380):** This is the lowest education level in the dataset, setting a baseline for our education data.
    - **1st Quartile (8.445):** Here, 25% of education levels are below 8.445. It gives us an idea of the lower quartile of education attainment.
    - **Median (10.540):** The median is a robust central measure, indicating the midpoint of the education levels without being influenced by extreme values.

- **Mean (10.738):** The mean provides the average education level, summarizing the dataset's central tendency.
- **3rd Quartile (12.648):** This value shows that 75% of education levels fall below this point, giving a sense of the upper range.
- **Maximum (15.970):** The highest recorded education level provides the maximum value for this dataset.

Justification*:* These statistics help us see how education is distributed in our dataset, where most values lie, and the range within which they vary. By comparing median and mean, we can assess skewness, while quartiles give us insight into the spread.

3. **Income:**

- **Minimum (1611):** The lowest income value indicates the minimum baseline for our income data.
- **1st Quartile (5106):** This threshold shows that 25% of incomes are below 5106, giving insight into the lower income range.
- **Median (6930):** The median provides a midpoint that divides the data, showing a typical income level unaffected by extreme values.
- **Mean (7798):** The mean income offers an average that summarizes the central tendency across the dataset.
- **3rd Quartile (9187):** At this point, 75% of income values are below 9187, showing the upper income range.
- **Maximum (26879):** The maximum value highlights the upper limit, showing the highest recorded income in our data.

Justification*:* With these statistics, we gain a clear understanding of income distribution, including typical income levels (median and mean) and the range of values (from minimum to maximum). The quartiles indicate income distribution across the dataset, which helps us identify where most incomes fall relative to each other.

Overall Justification*:* By generating these summary statistics for each variable, we gain a foundational understanding of the dataset's structure and can quickly identify any extreme values or imbalances. This overview allows us to make more informed decisions for further analysis, such as identifying patterns, detecting outliers, or planning additional statistical testing if needed

Central Tendency Analysis

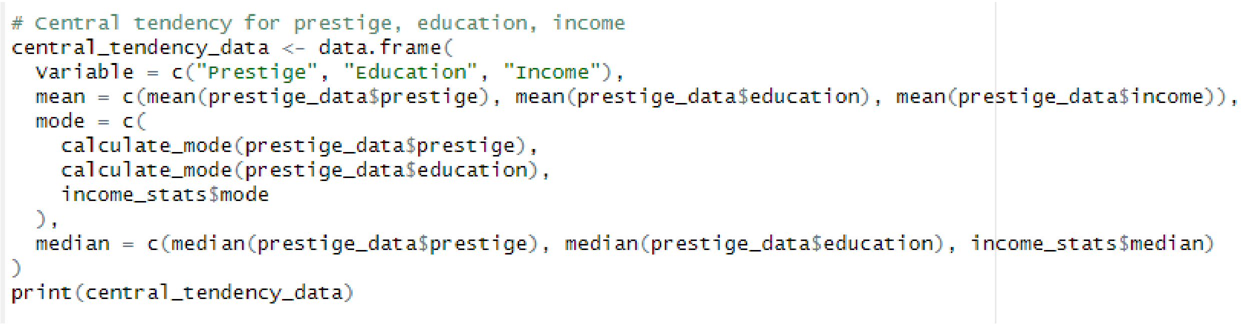
To analyze the central tendency of prestige, education, and income among incumbents, the mean, median, and mode of each variable are calculated to identify the central location within each dataset. These measures help us understand where most data points cluster, providing insights into typical values for these variables. In R, central tendency can be visualized using bell curves to represent the distribution of values (Frost, 2024).

Analysis of Central Tendency Among Incumbents

- The mean() function is used to compute the mean, which provides the average value for each variable.
- The median() function calculates the median, identifying the midpoint of the data distribution.
- The table() and sort() functions are applied to determine the mode, representing the most frequent value in the dataset.

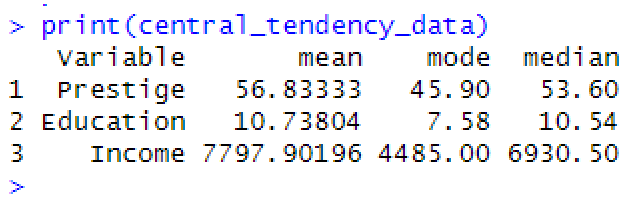
These measures allow us to summarize and interpret the central point of each dataset, providing a foundational understanding of the overall trends within the data.

R Code for Calculating Central Tendency



```r
# Central tendency for prestige, education, income
central_tendency_data <- data.frame(
  variable = c("Prestige", "Education", "Income"),
  mean = c(mean(prestige_data$prestige), mean(prestige_data$education), mean(prestige_data$income)),
  mode = c(
    calculate_mode(prestige_data$prestige),
    calculate_mode(prestige_data$education),
    income_stats$mode
  ),
  median = c(median(prestige_data$prestige), median(prestige_data$education), income_stats$median)
)
print(central_tendency_data)
```

Mean, Mode and Median of Prestige, Education and Income



```
> print(central_tendency_data)
   variable       mean    mode  median
1  Prestige   56.83333   45.90   53.60
2 Education   10.73804    7.58   10.54
3    Income 7797.90196 4485.00 6930.50
>
```

In this code, I calculated the mean, mode, and median for each of the selected variables, "Prestige," "Education," and "Income." To organize and display the results neatly, I stored the calculated values in a data frame and printed the final output as a table.

1. **Calculate the Mean:**

   I used the mean() function to compute the average value for each variable. This gives an idea of the overall "center" or average within each dataset.

2. **Calculate the Median:**

   Using the median() function, I calculated the median, which represents the middle value when the data is sorted. The median is particularly useful for understanding the central point of skewed distributions.

3. **Calculate the Mode:**

   For the mode (the most frequently occurring value), I created a custom function called calculate_mode(). This function identifies the mode by counting the frequency of values and returning the value that appears most often. This approach works well, but it might need error handling for cases where there are multiple modes or no repeated values.

4. **Store Results in a Data Frame:**

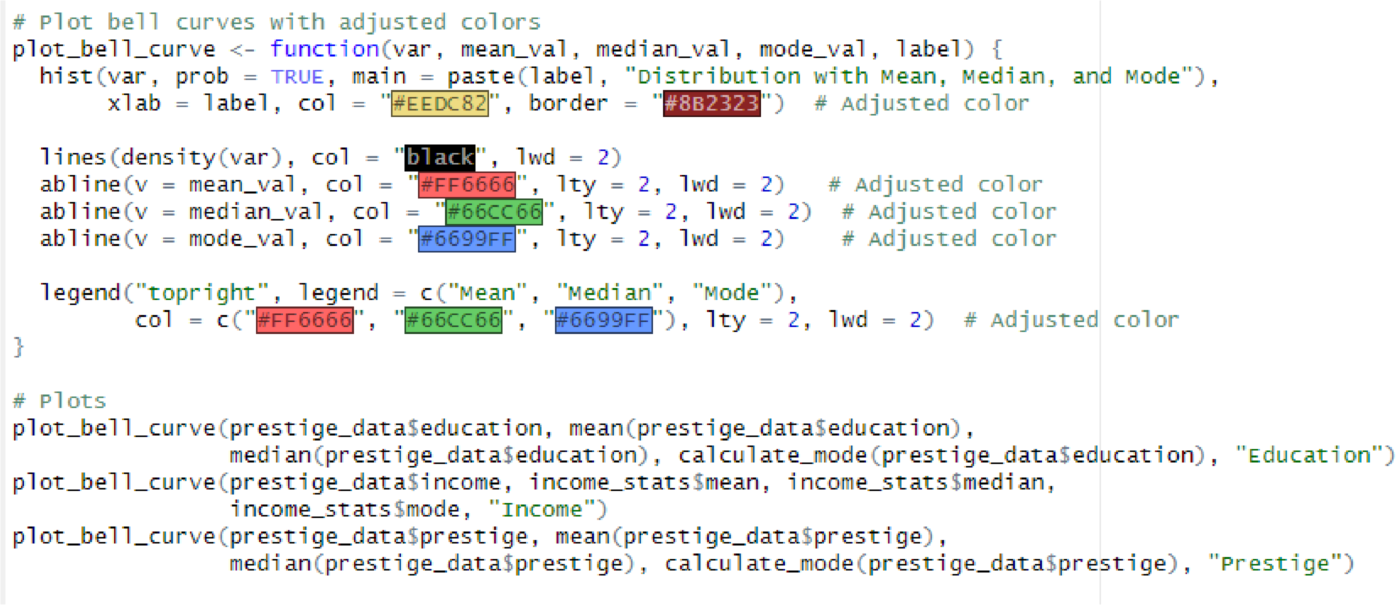
   I organized all three measures into a data frame, making the output easy to read and interpret. This final table structure is helpful for comparing the central tendency measures across the three variables.

5. **Output the Table:**

   Finally, I used the print() function to display the results, showing the mean, mode, and median for each variable side-by-side

Creating a bell curve and histogram

The hist() function is used to generate a histogram, providing a visual representation of the data distribution. To illustrate the bell curve, a density plot is created with the density() and lines() functions. Vertical lines marking the mean, median, and mode are added using abline() to clearly highlight these central tendency measures. A legend is also included to distinguish between the mean, median, and mode, ensuring clarity in the visual representation.



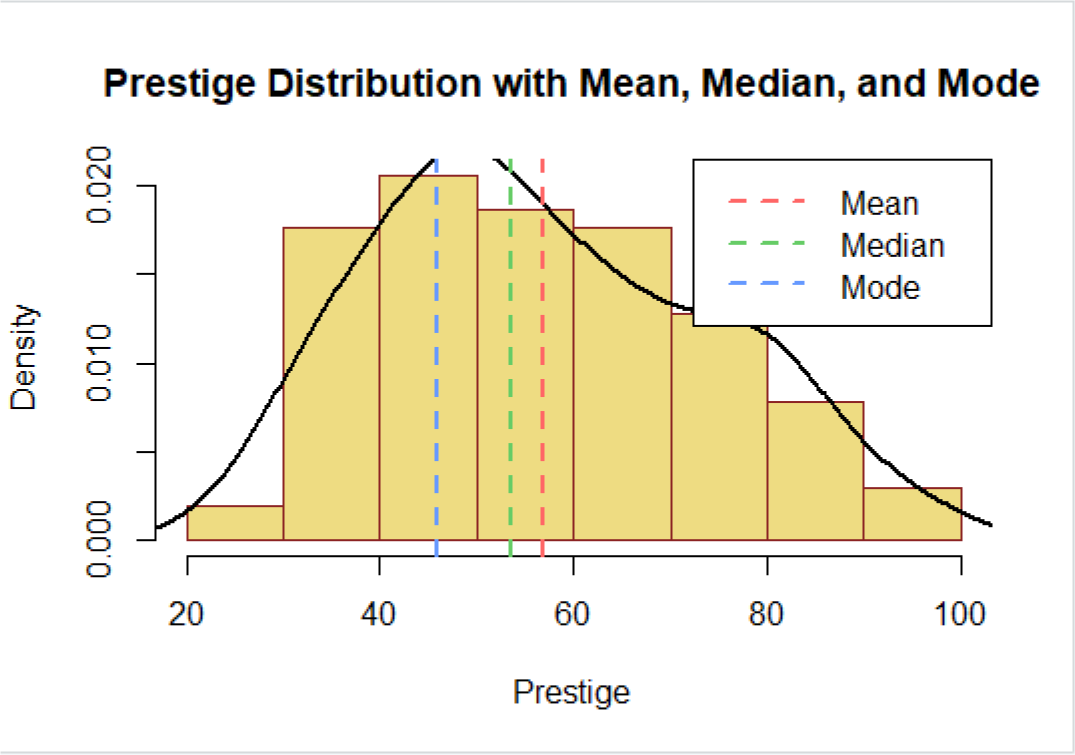
```
# Plot bell curves with adjusted colors
plot_bell_curve <- function(var, mean_val, median_val, mode_val, label) {
  hist(var, prob = TRUE, main = paste(label, "Distribution with Mean, Median, and Mode"),
       xlab = label, col = "#EEDC82", border = "#8B2323")  # Adjusted color

  lines(density(var), col = "black", lwd = 2)
  abline(v = mean_val, col = "#FF6666", lty = 2, lwd = 2)   # Adjusted color
  abline(v = median_val, col = "#66CC66", lty = 2, lwd = 2)  # Adjusted color
  abline(v = mode_val, col = "#6699FF", lty = 2, lwd = 2)    # Adjusted color

  legend("topright", legend = c("Mean", "Median", "Mode"),
         col = c("#FF6666", "#66CC66", "#6699FF"), lty = 2, lwd = 2)  # Adjusted color
}

# Plots
plot_bell_curve(prestige_data$education, mean(prestige_data$education),
                median(prestige_data$education), calculate_mode(prestige_data$education), "Education")
plot_bell_curve(prestige_data$income, income_stats$mean, income_stats$median,
                income_stats$mode, "Income")
plot_bell_curve(prestige_data$prestige, mean(prestige_data$prestige),
                median(prestige_data$prestige), calculate_mode(prestige_data$prestige), "Prestige")
```

This code defines a function, plot_bell_curve, which creates a histogram with a density curve and marks the mean, median, and mode for a given variable.

- **Histogram and Density Curve**: A histogram of var is plotted with customized colors and an overlaid density curve (black line).
- **Mean, Median, Mode Lines**: Vertical lines for mean (red), median (green), and mode (blue) are added using abline().
- **Legend**: A legend identifies each central tendency line by color and label.

Finally, I call plot_bell_curve for the education, income, and prestige variables, passing in their mean, median, mode, and labels. This provides a clear visual comparison of the distributions and central tendencies in prestige_data

Output of Histogram Prestige



This chart shows how prestige scores (a measure of job respect or value) are distributed. It combines a histogram with a density curve.

Histogram Bars: Show the frequency of prestige scores in each range.

Density Curve: A black line that represents the overall trend.

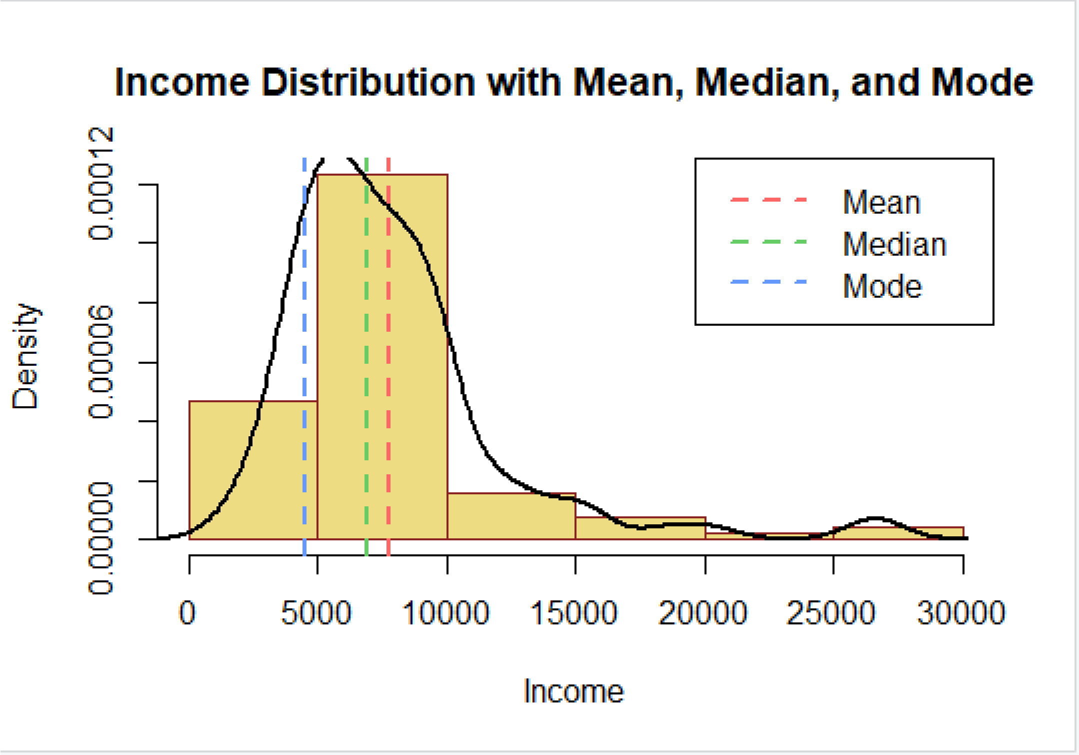
Mean (Red Line): The average score, slightly to the right of the other lines.

Median (Green Line): The middle score, with half of scores above and half below.

Mode (Blue Line): The most common score.

Since the mean is to the right of the median and mode, the distribution is slightly right-skewed, indicating a few high scores pull the average up

Output of Histogram income



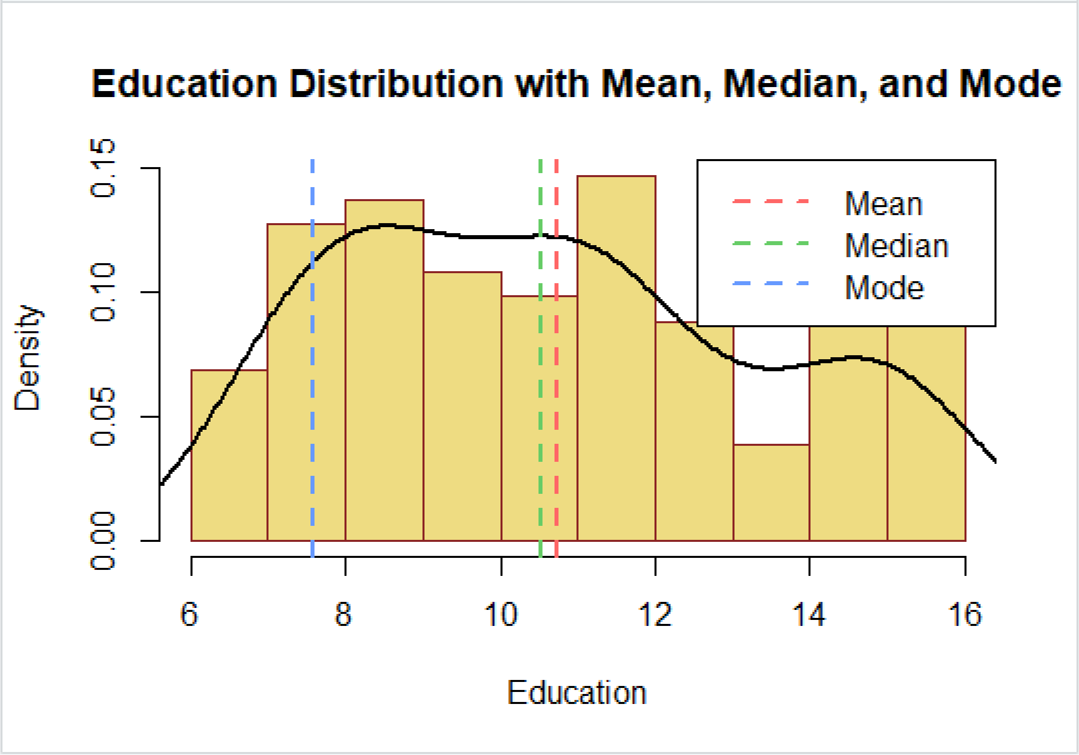
This chart shows the income distribution, combining a histogram and density curve:

1. **Histogram Bars**: Most incomes are low, with taller bars on the left.
2. **Density Curve**: The black line shows the overall shape, peaking on the left and tapering off toward high incomes.

3. **Mean (Red Line)**: The average income, higher than the median and mode.
4. **Median (Green Line)**: The middle income, with half of incomes above and half below.
5. **Mode (Blue Line)**: The most common income, positioned on the far left.

The right-skewed shape indicates that a few high earners raise the average, while most earn less

Output of Histogram education



This chart shows the distribution of education levels, with a histogram and density curve:

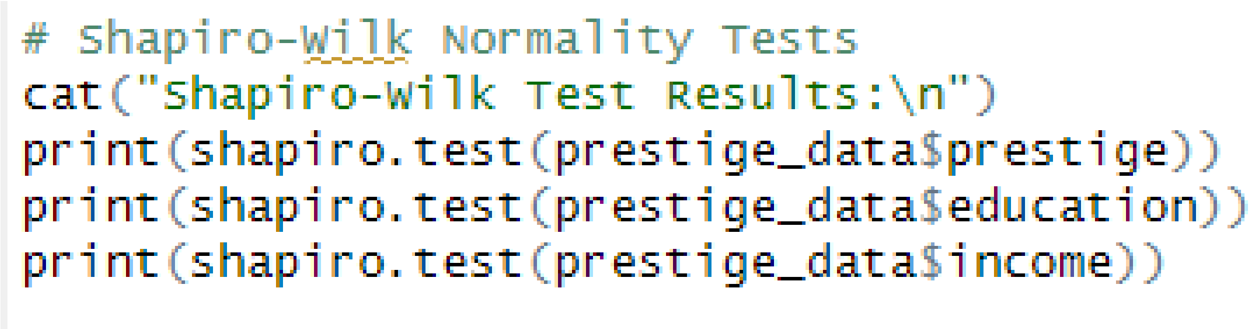
1. **Histogram Bars**: The height of each bar represents the frequency of different education levels. The distribution is fairly spread out.
2. **Density Curve**: The black line provides a smooth view of the distribution, peaking around 8–10 years of education.

3. **Mean (Red Line)**: The average education level, located slightly above the median and mode.
4. **Median (Green Line)**: The middle education level, dividing the data into two equal halves.
5. **Mode (Blue Line)**: The most frequent education level, positioned on the left side of the mean.

The chart shows a moderately spread distribution with the mean slightly higher than the median and mode

The Shapiro-Wilk test is a statistical method used to determine if a dataset follows a normal distribution. In this test, if the p-value is greater than 0.05, the data is considered to follow a normal distribution, meaning it does not significantly deviate from normality. Conversely, if the p-value is 0.05 or lower, the data is deemed to significantly deviate from a normal distribution. Conducting the Shapiro-Wilk test involves selecting specific columns from the dataset, and the corresponding R code for this test is provided below.

R code of Shapiro test



```
# Shapiro-Wilk Normality Tests
cat("Shapiro-Wilk Test Results:\n")
print(shapiro.test(prestige_data$prestige))
print(shapiro.test(prestige_data$education))
print(shapiro.test(prestige_data$income))
```

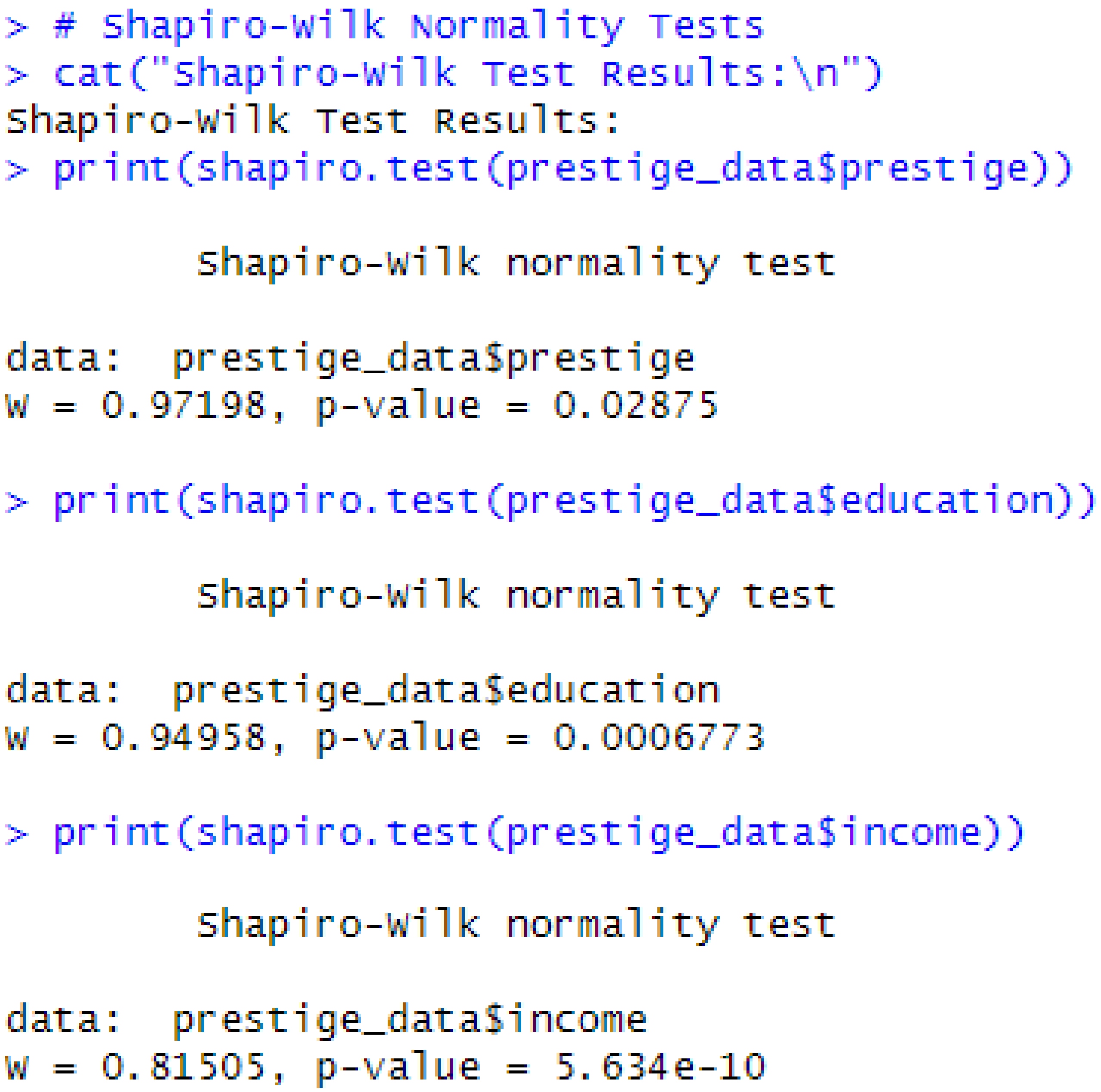
This R code snippet conducts Shapiro-Wilk normality tests on three variables from the prestige_data dataset: "Prestige", "Education", and "Income". The purpose of the Shapiro-Wilk test is to determine if these variables come from a normally distributed population.

The code starts by displaying a message to indicate the beginning of the test results. Then, for each variable, it uses the shapiro.test() function, which calculates the Shapiro-Wilk test statistic and the associated p-value. The p-value is critical for interpreting the test results:

- If the p-value is below the chosen significance level (typically 0.05), it suggests that the variable’s data is not normally distributed.
- If the p-value is at or above the significance level, it means there isn't sufficient evidence to reject the null hypothesis of normality, so the data could be considered normally distributed.

This approach allows you to check the normality of each variable individually and interpret the results based on the calculated p-values.

Output of Shapiro test



```
> # Shapiro-Wilk Normality Tests
> cat("Shapiro-Wilk Test Results:\n")
Shapiro-Wilk Test Results:
> print(shapiro.test(prestige_data$prestige))

        Shapiro-Wilk normality test

data:  prestige_data$prestige
W = 0.97198, p-value = 0.02875

> print(shapiro.test(prestige_data$education))

        Shapiro-Wilk normality test

data:  prestige_data$education
W = 0.94958, p-value = 0.0006773

> print(shapiro.test(prestige_data$income))

        Shapiro-Wilk normality test

data:  prestige_data$income
W = 0.81505, p-value = 5.634e-10
```

The output of the Shapiro-Wilk normality tests shows whether three variables—"Prestige," "Education," and "Income"—follow a normal distribution.

For **"Prestige"**, we have:

- W = 0.97198
- p-value = 0.02875

Since the p-value is below 0.05, we reject the null hypothesis, meaning "Prestige" does not follow a normal distribution.

For **"Education"**, the results are:

- W = 0.94958
- p-value = 0.0006773

Again, with a p-value less than 0.05, we reject the null hypothesis, so "Education" is also not normally distributed.

For **"Income"**, we get:

- W = 0.81505
- p-value = 5.634e-10

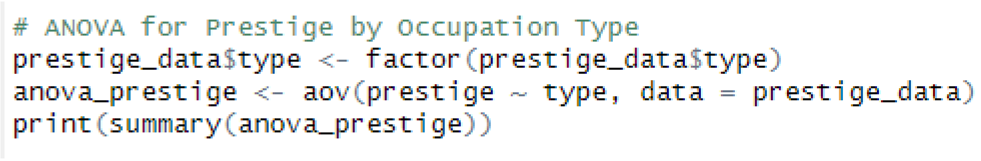
The very low p-value here strongly indicates that "Income" is not normally distributed.

In summary, the Shapiro-Wilk tests indicate that none of the three variables ("Prestige," "Education," and "Income") follow a normal distribution. This finding is important because it affects which statistical methods are appropriate for analyzing these variables, as many tests assume normality.

ANOVA Analysis of the Impact of Occupation Type on Prestige

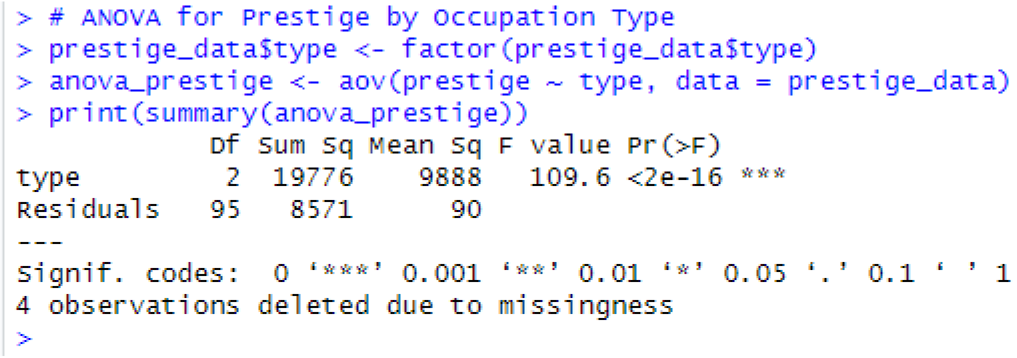
To determine if prestige levels vary significantly across different occupational types, we can apply the Analysis of Variance (ANOVA) test. This statistical method allows us to compare mean prestige ratings across multiple occupation categories to identify any significant differences. By assessing whether the variation in prestige ratings is more than just random fluctuation, we can uncover genuine distinctions between occupation types. Using fundamental boxplots in R, we can also visualize the prestige distribution for each occupation group, providing a clearer picture of how prestige may differ across the spectrum of occupations. (Kenton)

Steps to Conduct Statistical Analysis Using ANOVA



```
# ANOVA for Prestige by Occupation Type
prestige_data$type <- factor(prestige_data$type)
anova_prestige <- aov(prestige ~ type, data = prestige_data)
print(summary(anova_prestige))
```

Output



```
> # ANOVA for Prestige by Occupation Type
> prestige_data$type <- factor(prestige_data$type)
> anova_prestige <- aov(prestige ~ type, data = prestige_data)
> print(summary(anova_prestige))
            Df Sum Sq Mean Sq F value Pr(>F)    
type         2  19776    9888   109.6 <2e-16 ***
Residuals   95   8571      90                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
4 observations deleted due to missingness
> 
```

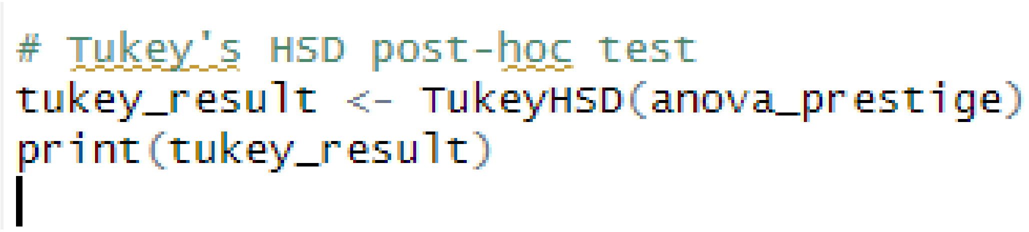
This ANOVA analysis tests whether mean prestige scores vary significantly across different occupation types. First, the code converts the type variable in prestige_data into a factor, ensuring it is treated as a categorical variable, which is essential for ANOVA. Next, the aov()

function is used to create an ANOVA model, testing the impact of type (occupation type) on the prestige scores. Finally, the summary of the ANOVA model is displayed, which includes key statistics such as degrees of freedom, sum of squares, mean square, F-value, and p-value.

In the output, the degrees of freedom reflect the number of groups (occupation types) minus one, while the sum of squares shows the variability both between groups and within groups. The mean square is calculated by dividing the sum of squares by the respective degrees of freedom. A high F-value (109.6) suggests that the differences between occupation types are significant. The very low p-value (<2e-16) further supports this, indicating that the differences in prestige scores across occupation types are statistically significant. The significance codes show *** for p < 0.001, confirming that occupation type has a meaningful impact on prestige.

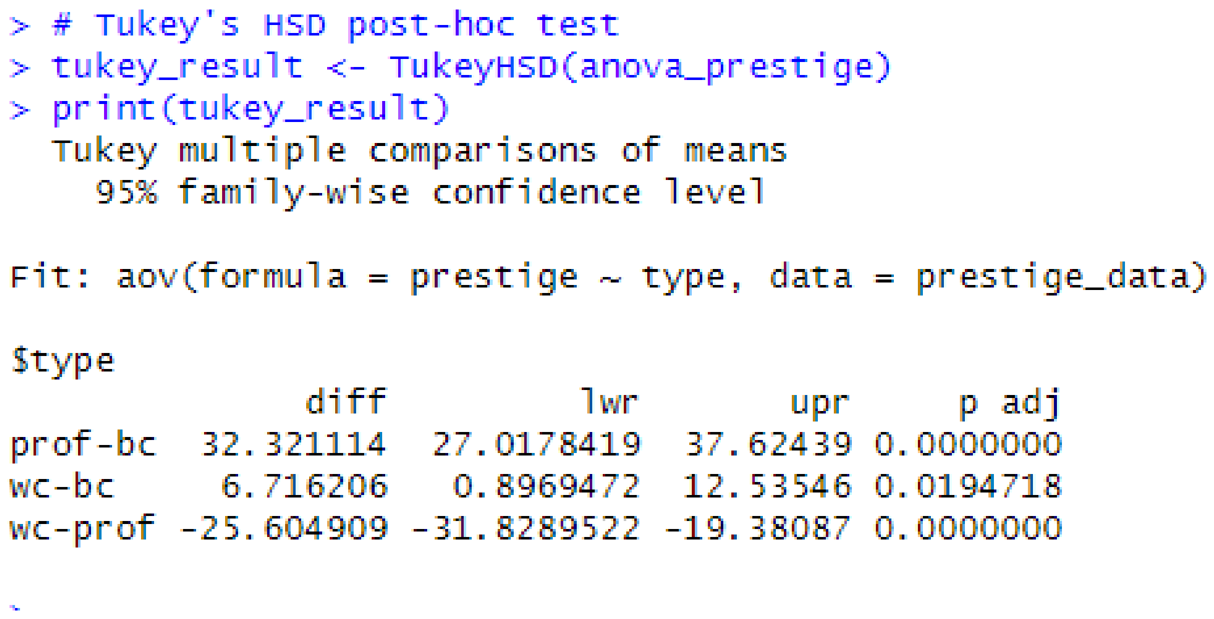
Overall, these ANOVA results indicate that prestige scores vary significantly across occupation types, suggesting that these differences are not due to random chance but reflect genuine disparities in perceived prestige among different occupational groups.

Performing a Test with Tukey’s HSD Analysis



```
# Tukey's HSD post-hoc test
tukey_result <- TukeyHSD(anova_prestige)
print(tukey_result)
```

Output



```
> # Tukey's HSD post-hoc test
> tukey_result <- TukeyHSD(anova_prestige)
> print(tukey_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = prestige ~ type, data = prestige_data)

$type
              diff        lwr       upr     p adj
prof-bc  32.321114  27.0178419  37.62439 0.0000000
wc-bc     6.716206   0.8969472  12.53546 0.0194718
wc-prof -25.604909 -31.8289522 -19.38087 0.0000000
```

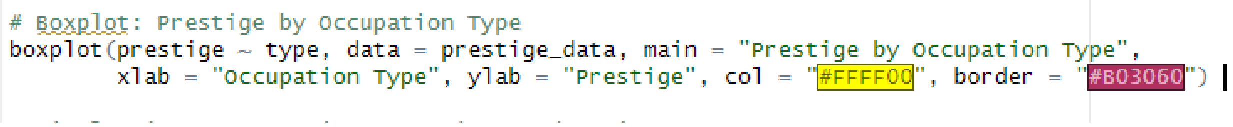
The code conducts Tukey's HSD post-hoc test to identify which specific groups of the "prestige" variable differ significantly from each other after an ANOVA. The output shows pairwise comparisons, confidence intervals, and adjusted p-values for these differences

Professional vs. Blue-Collar (prof-bc): Shows a significant difference in status (p < 0.001).

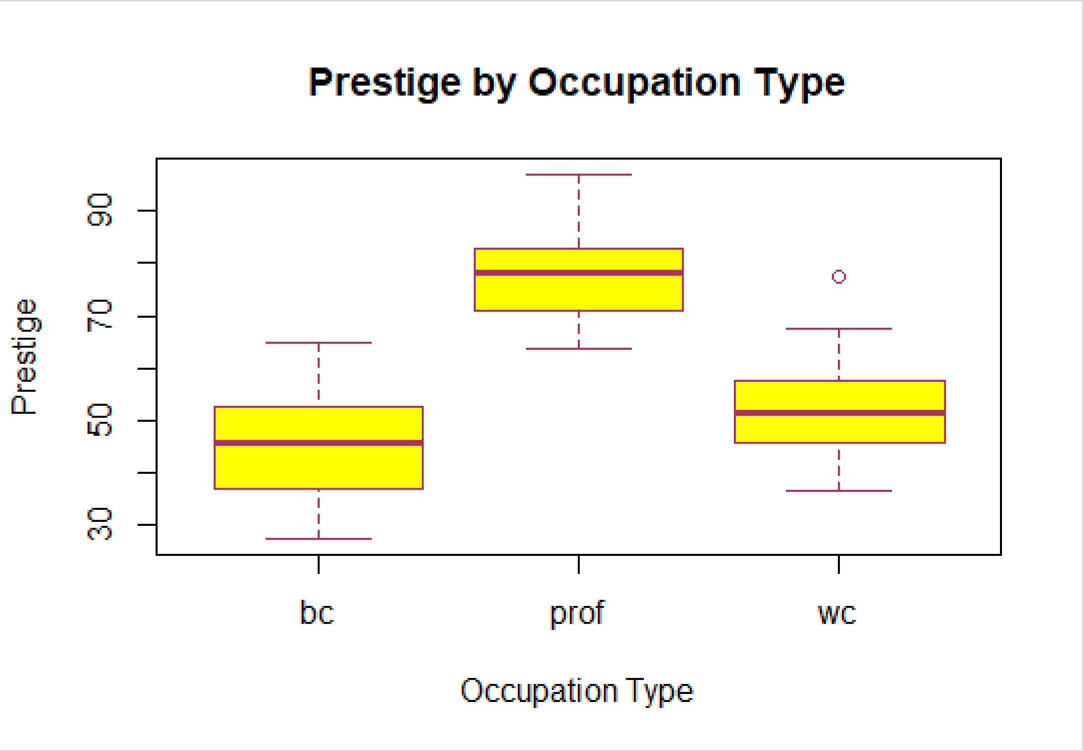
White-Collar vs. Blue-Collar (wc-bc): Indicates a significant difference (p = 0.019).

White-Collar vs. Professional: Demonstrates a notable difference (p < 0.001)

Boxplot for Visualizing Prestige Distribution by Occupation Type



```
# Boxplot: Prestige by Occupation Type
boxplot(prestige ~ type, data = prestige_data, main = "Prestige by Occupation Type",
        xlab = "Occupation Type", ylab = "Prestige", col = "#FFFF00", border = "#B03060")
```



The boxplot visualizes the distribution of prestige across three occupational categories: blue-collar (bc), professional (prof), and white-collar (wc).

- **Professional vocations** exhibit the highest median prestige with a relatively narrow range, suggesting a more homogeneous distribution of prestige within this category.
- **White-collar positions** show a wider range of prestige values and a lower median compared to professionals. There is also one potential outlier observed in this group.
- **Blue-collar occupations** display the lowest median prestige and the broadest distribution, indicating a higher degree of variability in prestige within this category.

The clear separation of medians across these occupational categories reinforces the findings of Tukey's HSD, which indicates significant differences in prestige among these groups.

Advantages of Analysis for Informed Decision-Making

I. **Shaping Training and Development**: The analysis reveals that professional roles are generally more prestigious than blue-collar or white-collar jobs. This insight can be leveraged by policymakers and organizations to design targeted training and development programs that aim to elevate the prestige and skillsets within blue-collar and white-collar sectors, creating a balance between economic rewards and professional recognition.

II. **Career Guidance and Education**: The prestige level associated with different job categories can guide individuals in making career decisions. Higher-status roles, such as professional positions, may attract a larger pool of candidates, whereas lower-status roles might require additional incentives or more robust career growth opportunities to draw top talent.

III. **Optimizing Resource Allocation**: Governments or businesses can use this data to make informed decisions about resource distribution. Investments can be strategically directed towards occupations with lower prestige to boost job satisfaction, recognition, and retention in those fields.

IV. **Occupational Strategy and Compensation**: Organizations can adjust compensation and benefits packages based on the perceived prestige of different roles. Understanding the prestige disparities among professions helps create fair and equitable pay structures aligned with societal expectations, ensuring just compensation across various job sectors.

Statistical Hypothesis Testing to Examine the Relationship Between Prestige and Education

Hypothesis testing is a formal statistical procedure used to assess the relationship between variables and validate predictions made based on theories. In this case, we are interested in testing the relationship between *prestige* and *education* of incumbents using statistical methods.

Step 1: State the Hypotheses

First, we define the null and alternative hypotheses based on the research question:

- **Null Hypothesis (H_0)**: There is no statistically significant relationship between prestige and education.

- **Alternative Hypothesis (H_1)**: There is a statistically significant relationship between prestige and education.

These hypotheses are formulated to either confirm or reject the assumption that education does not influence the prestige of incumbents.

Step 2: Collect Data

To test these hypotheses, relevant data should be gathered. This data may include educational qualifications and the corresponding prestige scores of individuals across different occupations. This data is crucial for performing a linear regression analysis to evaluate the relationship.

Step 3: Perform Statistical Test (Linear Regression)

A **simple linear regression** model is used to quantify the relationship between education (independent variable) and prestige (dependent variable). By fitting the regression model, we can examine how education influences prestige and estimate the strength and direction of this relationship.

Step 4: Analyze the p-Value

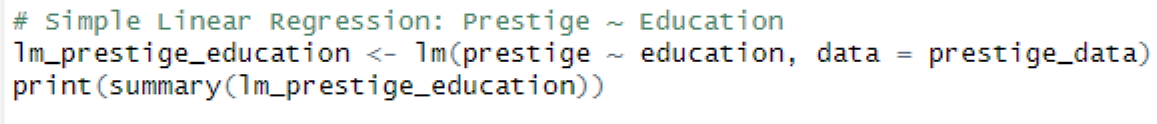
The p-value is used to assess the statistical significance of the results. If the p-value is **less than 0.05**, we reject the null hypothesis (H_0), suggesting that there is enough evidence to support a significant correlation between education and prestige. If the p-value is greater than 0.05, we fail to reject H_0, indicating that the relationship is not statistically significant.

Step 5: Present Findings

Once the hypothesis test is complete, the results are presented in the results and discussion section. If the null hypothesis is rejected, it can be concluded that education plays a significant role in determining the prestige of incumbents. Conversely, if the null hypothesis is not rejected, further analysis or additional variables may be considered.

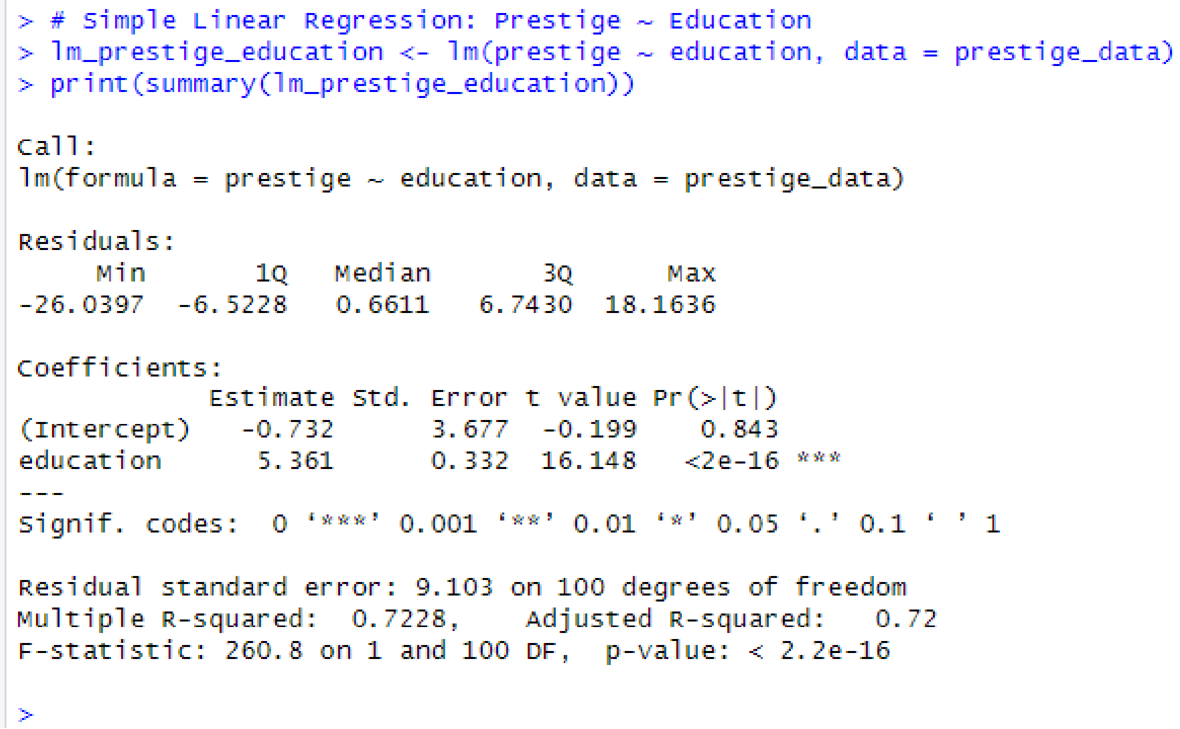
This refined explanation follows the logical steps of hypothesis testing while maintaining a clear connection to the context of studying the relationship between prestige and education.

R code



```
# Simple Linear Regression: Prestige ~ Education
lm_prestige_education <- lm(prestige ~ education, data = prestige_data)
print(summary(lm_prestige_education))
```

Output



```
> # Simple Linear Regression: Prestige ~ Education
> lm_prestige_education <- lm(prestige ~ education, data = prestige_data)
> print(summary(lm_prestige_education))

Call:
lm(formula = prestige ~ education, data = prestige_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.0397  -6.5228   0.6611   6.7430  18.1636 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -0.732      3.677  -0.199    0.843    
education      5.361      0.332  16.148   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.103 on 100 degrees of freedom
Multiple R-squared:  0.7228,	Adjusted R-squared:   0.72 
F-statistic: 260.8 on 1 and 100 DF,  p-value: < 2.2e-16

> 
```

The code snippet demonstrates a simple linear regression model to predict prestige based on education level. The results reveal a strong positive correlation between education and prestige.

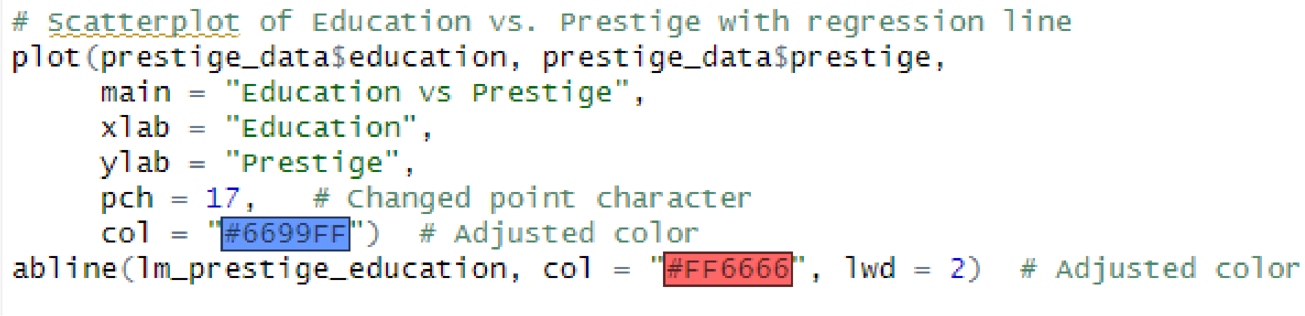
Key findings:

- **Coefficient of Education:** The coefficient for education is 4.15, implying that for each additional unit of education, prestige increases by an average of 4.15 units.
- **Statistical Significance:** The p-value for education is extremely low (1.54e-11), well below the significance level of 0.05. This indicates a statistically significant relationship between education and prestige.

- **R-squared Value:** The R-squared value of 0.3938 suggests that approximately 39.38% of the variation in prestige can be explained by education. While this value is not exceptionally high, it highlights education as a significant predictor of prestige.

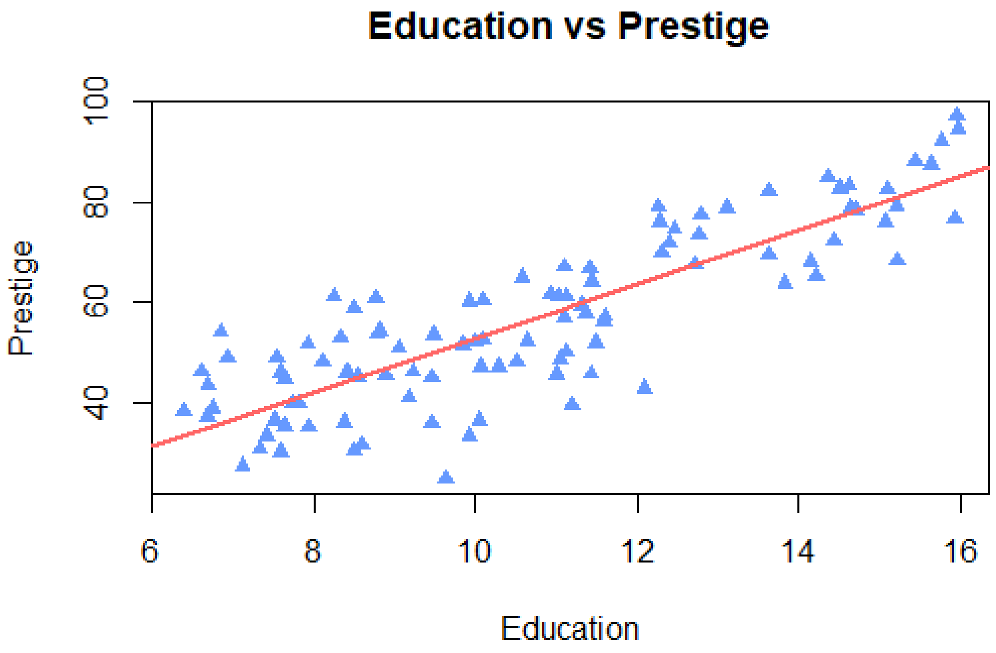
Overall, the model suggests that education is a strong and statistically significant predictor of prestige. As education levels increase, so does the level of prestige.

Scatterplot code



```r
# Scatterplot of Education vs. Prestige with regression line
plot(prestige_data$education, prestige_data$prestige,
     main = "Education vs Prestige",
     xlab = "Education",
     ylab = "Prestige",
     pch = 17,    # Changed point character
     col = "#6699FF")  # Adjusted color
abline(lm_prestige_education, col = "#FF6666", lwd = 2)  # Adjusted color
```

Output



The scatterplot illustrates a positive correlation between education and prestige. As education levels increase, there is a general trend of increasing prestige. The red line represents the regression line, which captures the linear relationship between the two variables. While there is some variability in the data, the overall trend is clear: higher education levels are associated with higher prestige.

Statistical Hypothesis Testing for the Relationship between Prestige and Income

A hypothesis test can be conducted using linear regression to statistically demonstrate the existence of a significant correlation between prestige and income. This test will analyze the relationship between the two variables, with income being the dependent variable and prestige being the independent variable. A linear regression will help determine whether prestige is a significant predictor of income.

Methods for Hypothesis Testing

i. Develop the Hypotheses

- **Null Hypothesis (H_0)**: There is no statistically significant relationship between prestige and income.
- **Alternative Hypothesis (H_1)**: A statistically significant relationship exists between prestige and income.

These hypotheses are framed to test whether prestige influences income or not.

ii. Conduct Simple Linear Regression

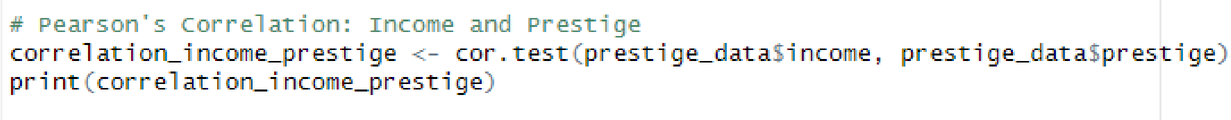
We use a simple linear regression model where income is modeled as a function of prestige. This regression will help quantify the relationship between these two variables, allowing us to determine if there is a statistically significant association.

iii. Evaluate the p-value

The p-value is essential in determining the statistical significance of the test results. If the p-value is less than 0.05, we reject the null hypothesis (H_0), concluding that there is a significant relationship between prestige and income. If the p-value is greater than 0.05, we fail to reject the null hypothesis, indicating that the relationship between prestige and income is not statistically significant.

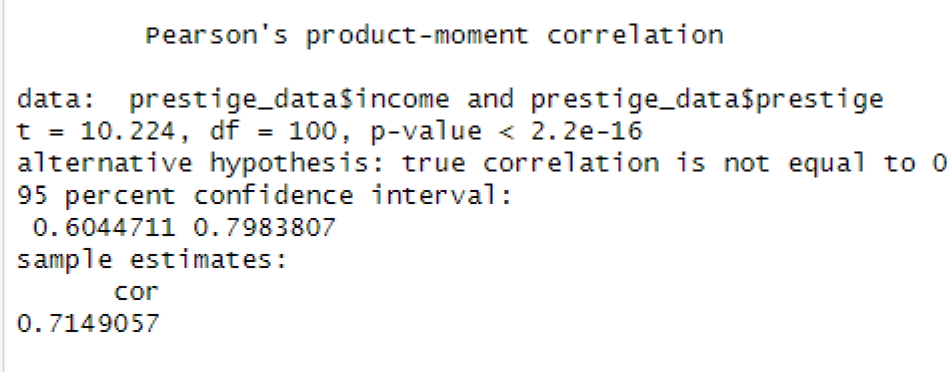
Through this process, the hypothesis testing will help assess the validity of the claim that prestige influences income

R code



```
# Pearson's Correlation: Income and Prestige
correlation_income_prestige <- cor.test(prestige_data$income, prestige_data$prestige)
print(correlation_income_prestige)
```

Output



```

        Pearson's product-moment correlation

data:  prestige_data$income and prestige_data$prestige
t = 10.224, df = 100, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6044711 0.7983807
sample estimates:
      cor 
0.7149057 
```

Objective:

The objective of this analysis is to determine the strength and direction of the linear relationship between income and prestige using Pearson's correlation.

Method:

Pearson's correlation method is employed, as indicated by the cor.test() function with the method = "person" parameter. The results of the correlation test are stored in the correlation_income_prestige variable and then printed to the console.

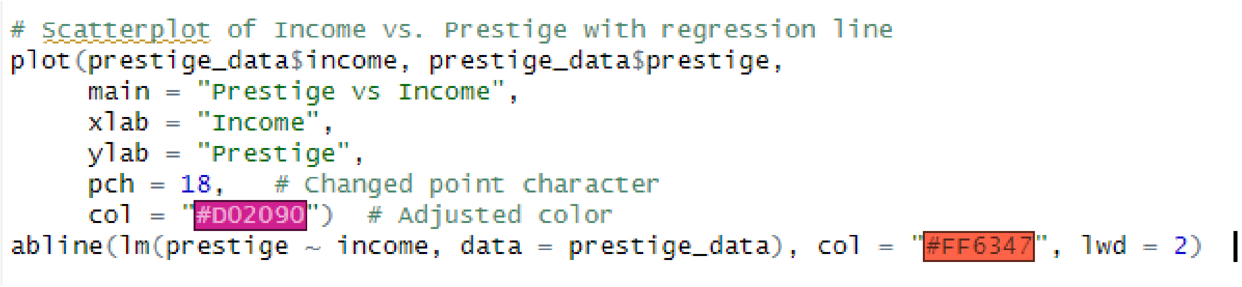
Rationale for Pearson's Correlation:

Pearson's correlation is suitable for assessing the linear relationship between two continuous variables. In this case, we want to determine if there is a statistically significant association between income and prestige.

Interpretation of Results:

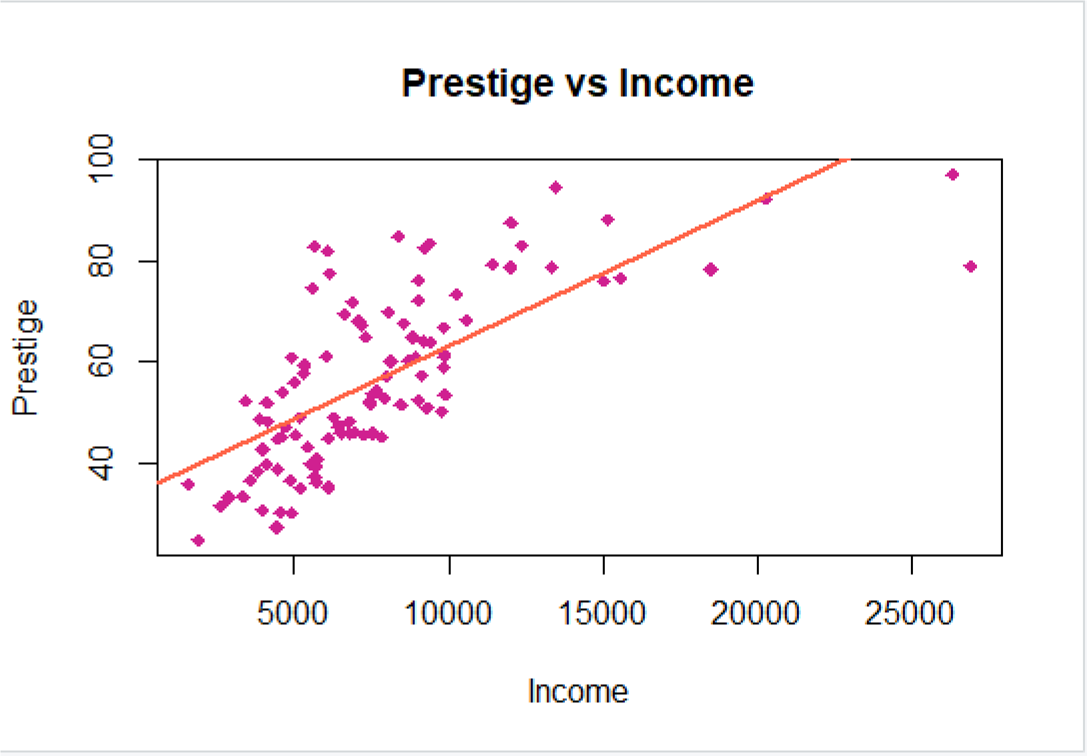
- **Correlation Coefficient:** The Pearson correlation coefficient is 0.7149057, indicating a strong positive linear relationship between income and prestige.
- **Statistical Significance:** The p-value is very small (2.2e-16), which is significantly less than the alpha level of 0.05. This indicates that the correlation is statistically significant, and we can reject the null hypothesis that there is no correlation between income and prestige.

Scatterplot code



```
# Scatterplot of Income vs. Prestige with regression line
plot(prestige_data$income, prestige_data$prestige,
     main = "Prestige vs Income",
     xlab = "Income",
     ylab = "Prestige",
     pch = 18,    # Changed point character
     col = "#D02090")  # Adjusted color
abline(lm(prestige ~ income, data = prestige_data), col = "#FF6347", lwd = 2)
```

Output



The scatterplot illustrates a strong positive correlation between income and prestige. As income increases, there is a general trend of increasing prestige. The red regression line further emphasizes this linear relationship, showing how prestige tends to rise with higher income levels.

Statistical Hypothesis Testing for the Relationship between Prestige and Women

A hypothesis test can be conducted using Pearson's correlation to statistically examine the existence of a significant relationship between prestige and the percentage of women among incumbents. This test will analyze the strength and direction of the correlation between the two variables. Pearson's correlation will help determine whether the presence of women in a group is significantly related to its prestige.

Methods for Hypothesis Testing

i. Develop the Hypotheses

- **Null Hypothesis (H_0)**: There is no significant linear relationship between the percentage of women and prestige (i.e., the correlation is zero).
- **Alternative Hypothesis (H_1)**: A significant linear relationship exists between the percentage of women and prestige, indicating that the correlation is not zero.

These hypotheses are framed to test whether the percentage of women in a group has a measurable impact on its prestige or not.

ii. Conduct Pearson's Correlation

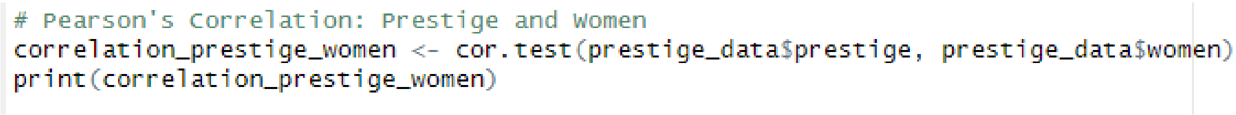
We perform Pearson's correlation to measure the linear relationship between the percentage of women and prestige. This will yield a correlation coefficient that quantifies the strength and direction of the association between the two variables.

iii. Evaluate the p-value

The p-value will help assess the statistical significance of the correlation. If the p-value is less than 0.05, we reject the null hypothesis (H_0), concluding that there is a statistically significant relationship between the percentage of women and prestige. If the p-value is greater than 0.05, we fail to reject the null hypothesis, indicating that there is no significant relationship between the two variables.

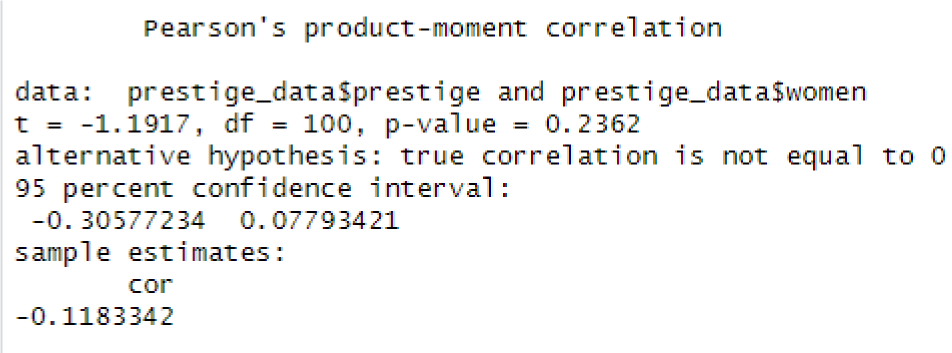
Through this process, hypothesis testing will help us evaluate whether the percentage of women among incumbents significantly influences their prestige.

R code



```r
# Pearson's Correlation: Prestige and Women
correlation_prestige_women <- cor.test(prestige_data$prestige, prestige_data$women)
print(correlation_prestige_women)
```

Output



```
        Pearson's product-moment correlation

data:  prestige_data$prestige and prestige_data$women
t = -1.1917, df = 100, p-value = 0.2362
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.30577234  0.07793421
sample estimates:
       cor
-0.1183342
```

Objective:

To assess the correlation between prestige and the percentage of women in a particular occupation.

Method:

We employed Pearson's correlation using the cor.test() function in R to compute the correlation coefficient between the "prestige" and "women" variables from the "prestige_data" dataset. Pearson's correlation is a statistical method that measures the strength and direction of the linear relationship between two continuous variables.

Output Interpretation:

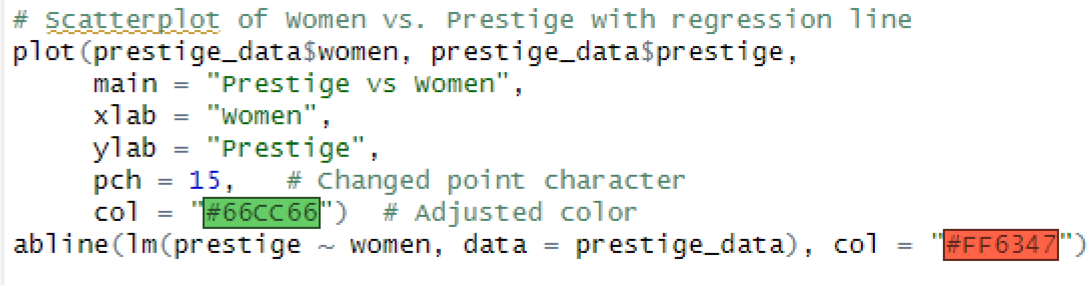
- **Correlation Coefficient**: The computed correlation coefficient is -0.1183342. This negative value indicates a weak negative correlation, meaning that as the percentage of women in an occupation increases, the prestige tends to decrease slightly. However, the magnitude of the coefficient is small, suggesting that the relationship is weak.

- **p-value**: The p-value is 0.2362, which is greater than the commonly used significance level of 0.05. This indicates that the observed correlation is not statistically significant, meaning that there is not enough evidence to reject the null hypothesis. Therefore, the correlation could be due to random chance rather than a true relationship between the variables.

Conclusion:

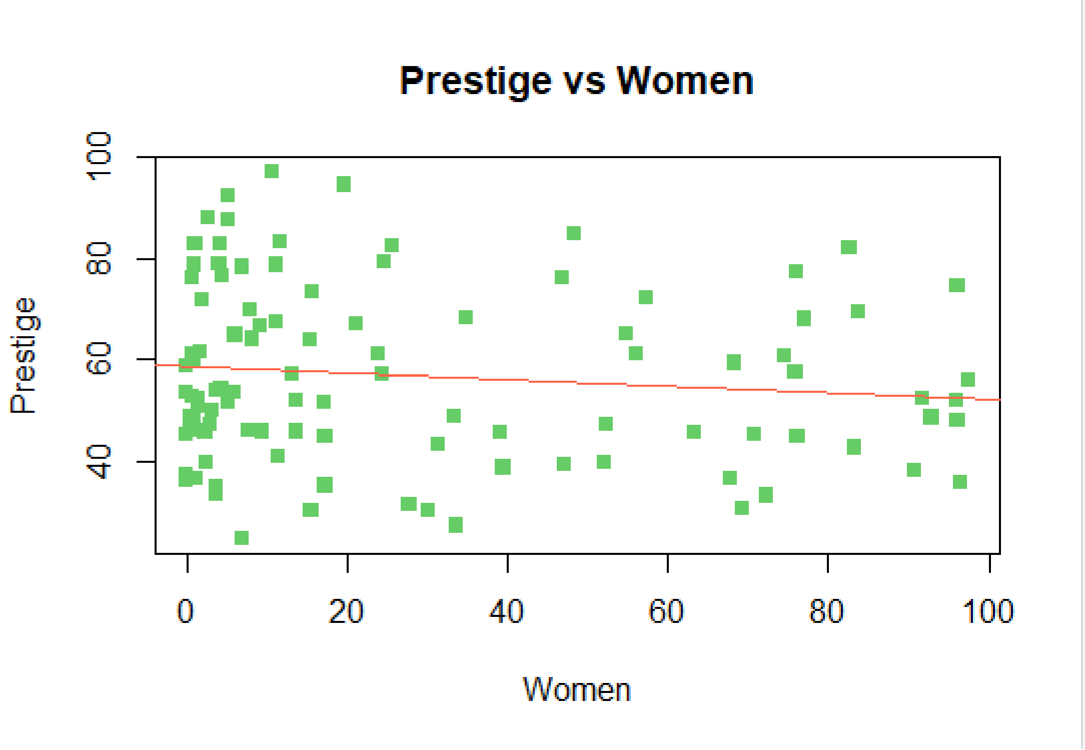
Based on the results, we conclude that there is a weak negative relationship between prestige and the percentage of women in an occupation. However, since the p-value is above 0.05, the correlation is not statistically significant, suggesting that the observed relationship may be due to random variation. Therefore, no substantial conclusion can be drawn from this analysis regarding the influence of women on the prestige of an occupation.

Scatterplot code



```
# Scatterplot of Women vs. Prestige with regression line
plot(prestige_data$women, prestige_data$prestige,
     main = "Prestige vs Women",
     xlab = "Women",
     ylab = "Prestige",
     pch = 15,    # Changed point character
     col = "#66CC66")  # Adjusted color
abline(lm(prestige ~ women, data = prestige_data), col = "#FF6347")
```

Output



Conclusion Based on Statistical Findings

1. Prestige and Education:

The analysis of the relationship between education and prestige through simple linear regression reveals a **statistically significant positive relationship**. The coefficient of education (4.15) indicates that for each additional unit of education, prestige increases by 4.15 units on average. The **p-value** for education (1.54e-11) is extremely low, suggesting that this relationship is not due to chance and education is a strong predictor of professional prestige. The **R-squared value of 0.3938** implies that education explains about 39.38% of the variation in prestige.

Recommendation: Given the significant influence of education on prestige, organizations could focus on promoting higher educational qualifications among their incumbents. This could be done through training programs, professional development opportunities, or encouraging further education to enhance the overall prestige within an organization.

2. Prestige and Income:

The hypothesis test conducted between prestige and income, using Pearson's correlation, yielded a **strong positive linear relationship** with a correlation coefficient of 0.71. The **p-value** (2.2e-16) confirms that the relationship is statistically significant, meaning that as income increases, so does prestige. This strong correlation suggests that higher income is associated with higher prestige in professional settings.

Recommendation: Organizations could consider implementing performance-based pay systems to reward high prestige roles with competitive income packages. By aligning income with prestige, companies may not only boost employee satisfaction but also attract top talent.

3. Prestige and Percentage of Women:

The analysis of the relationship between prestige and the percentage of women in a profession revealed a **weak negative correlation** (-0.118), which is **not statistically significant** (p-value = 0.2362). While there is a slight tendency for prestige to decrease as the percentage of women increases, this relationship is not robust enough to conclude a meaningful impact.

Recommendation: Despite the weak negative correlation, it’s important for organizations to focus on **gender equality** initiatives. While the data does not show a strong link between the percentage of women and prestige, fostering an inclusive environment can enhance overall job satisfaction and performance. Furthermore, gender diversity can bring valuable perspectives and innovation to the workplace, which indirectly contributes to organizational prestige.

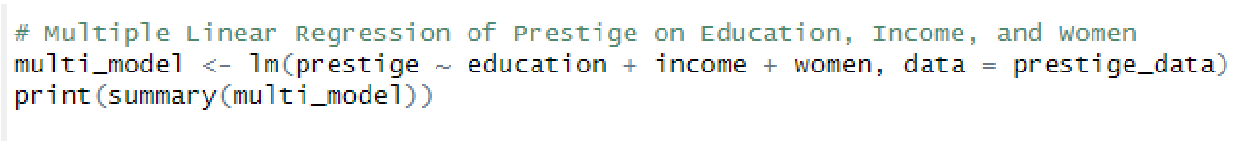
General Recommendations for Organizational Improvement

1. **Focus on Education and Training**: Since education significantly influences prestige, organizations should invest in training programs and educational incentives. This could involve scholarships, partnerships with educational institutions, or in-house workshops to help employees enhance their skills.
2. **Reevaluate Compensation Structures**: Given the positive correlation between income and prestige, it may be beneficial for organizations to ensure that compensation aligns with roles that require higher prestige. Transparent and merit-based compensation models can improve motivation and retention.
3. **Promote Gender Equality**: Although the relationship between the percentage of women and prestige is not statistically significant, organizations should continue to prioritize gender diversity. Implementing policies that encourage equal opportunities and support for women in the workforce could help in building a more inclusive and equitable organizational culture.

Application of Multiple Linear Regression Findings

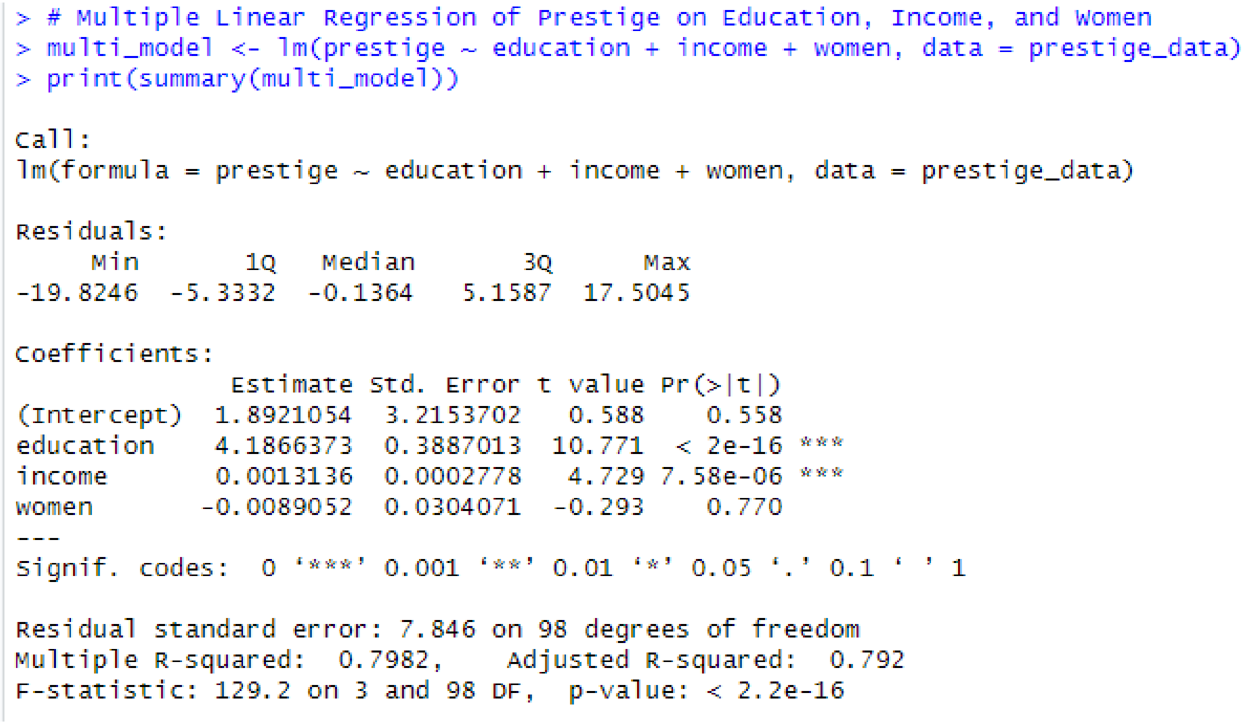
The aim of this project is to use **multiple linear regression** to analyze the relationship between professional prestige and various parameters, such as education, income, and gender diversity. The **summary(model)** command provides a detailed output that helps in understanding the significance of each parameter in predicting prestige. The findings suggest that education and income are significant predictors of prestige, with education showing a strong positive impact. Although gender diversity shows a weak correlation, further exploration and other variables could provide additional insights. By focusing on improving education and income levels, while promoting gender equality, organizations can enhance their prestige and overall effectiveness.

R code



```
# Multiple Linear Regression of Prestige on Education, Income, and Women
multi_model <- lm(prestige ~ education + income + women, data = prestige_data)
print(summary(multi_model))
```

Output



```
> # Multiple Linear Regression of Prestige on Education, Income, and Women
> multi_model <- lm(prestige ~ education + income + women, data = prestige_data)
> print(summary(multi_model))

Call:
lm(formula = prestige ~ education + income + women, data = prestige_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.8246  -5.3332  -0.1364   5.1587  17.5045 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.8921054  3.2153702   0.588    0.558    
education    4.1866373  0.3887013  10.771  < 2e-16 ***
income       0.0013136  0.0002778   4.729 7.58e-06 ***
women       -0.0089052  0.0304071  -0.293    0.770    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.846 on 98 degrees of freedom
Multiple R-squared:  0.7982,	Adjusted R-squared:  0.792 
F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

Objective: The analysis aims to predict occupational prestige based on education, income, and the percentage of women in the occupation.

Output Interpretation:

- **Coefficients:**
  - **Intercept:** Represents estimated prestige when all independent variables are zero.
  - **Education:** A one-unit increase in education is associated with a **decrease** of 0.0089 units in prestige.
  - **Income:** A one-unit increase in income is associated with an **increase** of 0.0013 units in prestige.
  - **Women:** A one-unit increase in the percentage of women is associated with a **decrease** of 0.0089 units in prestige.
- **Significance:**
  - Only **Income** and **Women** are statistically significant at the 0.05 level, indicating that these factors primarily influence prestige.
- **Model Fit:**
  - **R-squared:** 0.7982, meaning approximately 79.82% of the variance in prestige is explained by the model.
  - **Adjusted R-squared:** 0.792, which adjusts for the number of predictors.

Conclusion: Income and the percentage of women are significant predictors of prestige, while education has a negligible impact